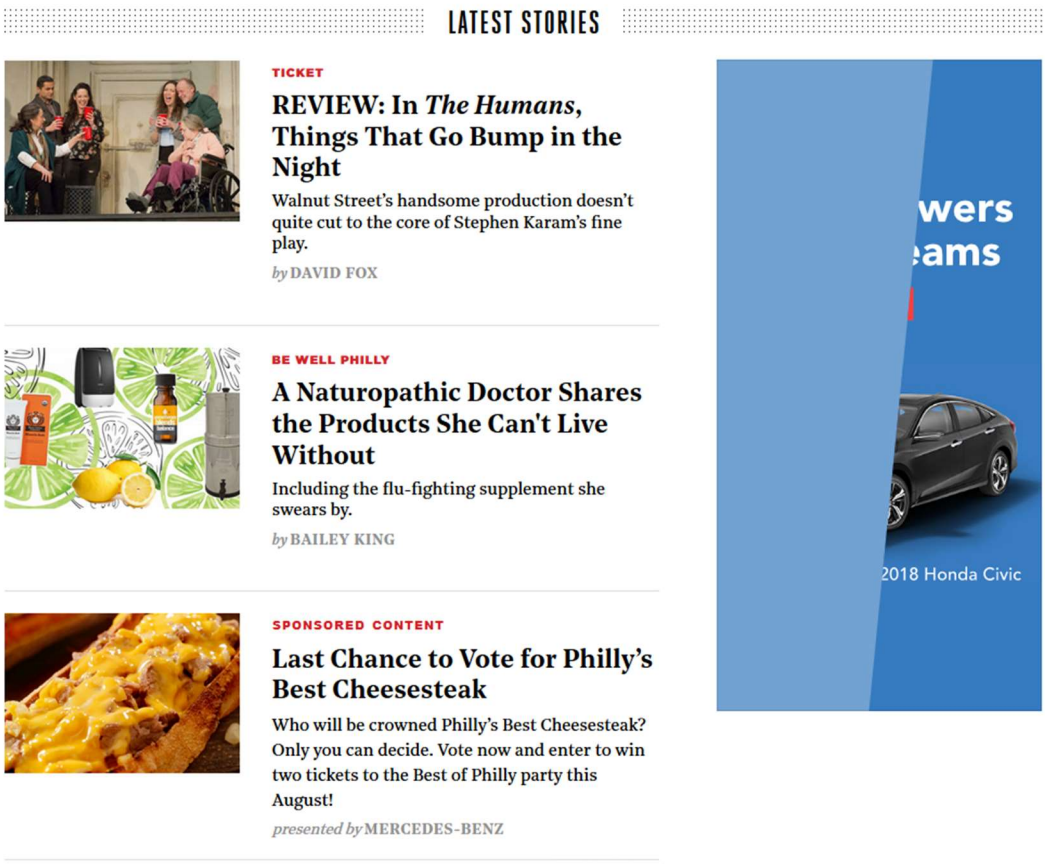
MATE-T580 Practical Data Science using R
Assignment 2
Extracting Data from the Web

Your boss has asked you to write an R script to keep track of **LATEST STORIES** published on Philadelphia Magazine (www.phillymag.com). Here's a description of what the script should do:

- The only input to the script is the web address of the magazine (i.e. www.phillymag.com)
- The script will fetch the webpage, look for articles listed under the **LATEST STORIES** section (see screenshot below), and parse the html content of the web page as such to produce a data frame with the following features:
  - Each row of the data frame is an article listed under **LATEST STORIES**.
  - The data frame has 5 columns containing information on: Date article published, title of article, article category (red tag), author of article, and http address of article.
- The script should also download each article listed in the data frame and save its content in the from of a .txt file while (as much as possible) stripping away any text that is not part of the article (e.g. webpage links, advertisements, etc.). So if the web page contained 12 articles on the date the script was run, 12 different files should be saved in a folder, one for each article.

Note: This problem requires use of XML parsing functionality (*XML* library) and some 'gentle' use of regular expressions (either under R BASE or the *stringr* library). Also note that since the webpage is very well structured (in terms of adoption of well thought of hierarchical xml tagging structure), I expect that there will be many different ways to extract the same piece information. This should make your job easier!

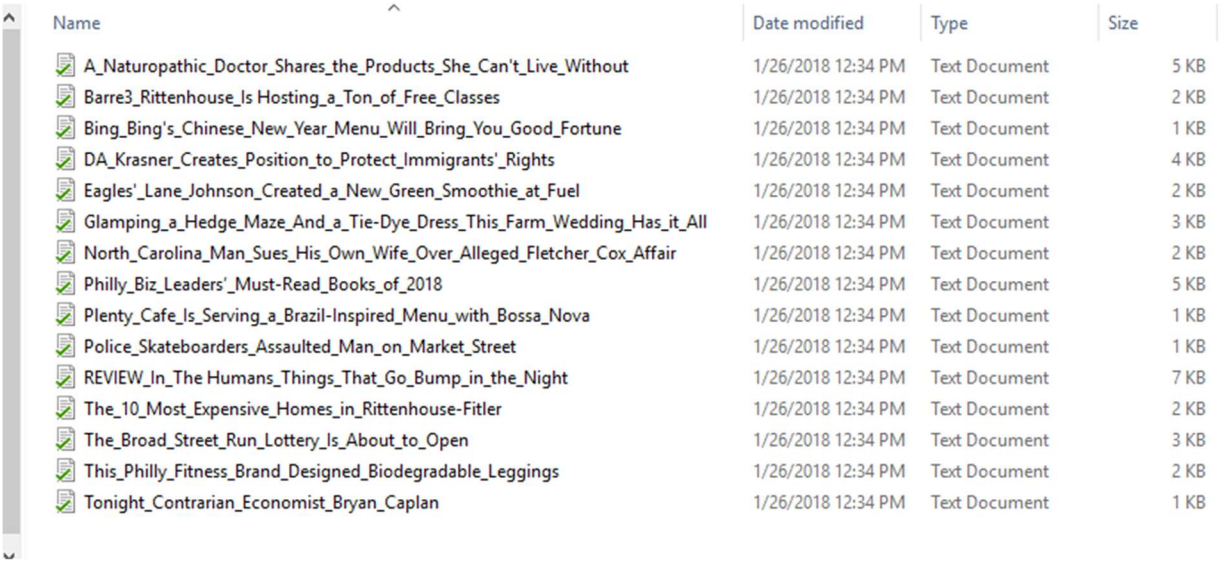
Screenshot from phillymag.com showing part of the LATEST STORIES section. Note: you may ignore articles with the SPONSORED CONTENT tag.



Screenshot from the RStudio console showing the first 5 rows of the data fame that contains information on the articles

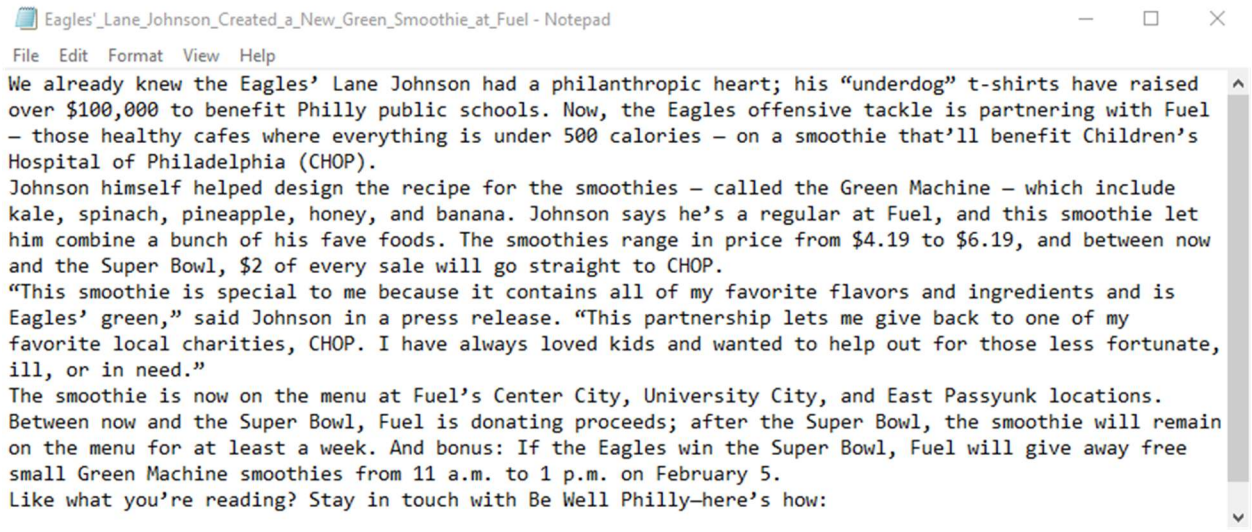
```
Console ~/
> top_n(Phillymag_articles,5)
Selecting by Link
# A tibble: 5 x 5
        Date                                                                  Title             Category        Author
      <date>                                                                  <chr>                <chr>         <chr>
1 2018-01-26                  REVIEW: In The Humans, Things That Go Bump in the Night             Ticket     David Fox
2 2018-01-25 Glamping, a Hedge Maze, And a Tie-Dye Dress: This Farm Wedding Has it All Philadelphia Wedding   Bailey King
3 2018-01-25                      Police: Skateboarders Assaulted Man on Market Street               News Joe Trinacria
4 2018-01-25                         The 10 Most Expensive Homes in Rittenhouse-Fitler           Property   Sandy Smith
5 2018-01-25                                  Tonight: Contrarian Economist Bryan Caplan             Ticket  Patrick Rapa
# ... with 1 more variables: Link <chr>
> Phillymag_articles$Link[1:5]
[1] "http://www.phillymag.com/ticket/2018/01/26/review-humans-things-go-bump-night/"
[2] "http://www.phillymag.com/be-well-philly/2018/01/26/tara-nayak-favorite-products/"
[3] "http://www.phillymag.com/philadelphia-wedding/2018/01/25/hippie-farm-wedding/"
[4] "http://www.phillymag.com/news/2018/01/25/fletcher-cox-girlfriend-lawsuit-catherine-cuesta-joshua-jeffords/"
[5] "http://www.phillymag.com/news/2018/01/25/skateboard-assault-market-street/"
>
```

Screenshot from windows explorer showing the articles saved as .txt files.



Name	Date modified	Type	Size
A_Naturopathic_Doctor_Shares_the_Products_She_Can't_Live_Without	1/26/2018 12:34 PM	Text Document	5 KB
Barre3_Rittenhouse_Is Hosting_a_Ton_of_Free_Classes	1/26/2018 12:34 PM	Text Document	2 KB
Bing_Bing's_Chinese_New_Year_Menu_Will_Bring_You_Good_Fortune	1/26/2018 12:34 PM	Text Document	1 KB
DA_Krasner_Creates_Position_to_Protect_Immigrants'_Rights	1/26/2018 12:34 PM	Text Document	4 KB
Eagles'_Lane_Johnson_Created_a_New_Green_Smoothie_at_Fuel	1/26/2018 12:34 PM	Text Document	2 KB
Glamping_a_Hedge_Maze_And_a_Tie-Dye_Dress_This_Farm_Wedding_Has_it_All	1/26/2018 12:34 PM	Text Document	3 KB
North_Carolina_Man_Sues_His_Own_Wife_Over_Alleged_Fletcher_Cox_Affair	1/26/2018 12:34 PM	Text Document	2 KB
Philly_Biz_Leaders'_Must-Read_Books_of_2018	1/26/2018 12:34 PM	Text Document	5 KB
Plenty_Cafe_Is_Serving_a_Brazil-Inspired_Menu_with_Bossa_Nova	1/26/2018 12:34 PM	Text Document	1 KB
Police_Skateboarders_Assaulted_Man_on_Market_Street	1/26/2018 12:34 PM	Text Document	1 KB
REVIEW_In_The Humans_Things_That_Go_Bump_in_the_Night	1/26/2018 12:34 PM	Text Document	7 KB
The_10_Most_Expensive_Homes_in_Rittenhouse-Fitler	1/26/2018 12:34 PM	Text Document	2 KB
The_Broad_Street_Run_Lottery_Is_About_to_Open	1/26/2018 12:34 PM	Text Document	3 KB
This_Philly_Fitness_Brand_Designed_Biodegradable_Leggings	1/26/2018 12:34 PM	Text Document	2 KB
Tonight_Contrarian_Economist_Bryan_Caplan	1/26/2018 12:34 PM	Text Document	1 KB

Sample article displayed in Notepad



Eagles'_Lane_Johnson_Created_a_New_Green_Smoothie_at_Fuel - Notepad

File Edit Format View Help

We already knew the Eagles' Lane Johnson had a philanthropic heart; his "underdog" t-shirts have raised over $100,000 to benefit Philly public schools. Now, the Eagles offensive tackle is partnering with Fuel – those healthy cafes where everything is under 500 calories – on a smoothie that'll benefit Children's Hospital of Philadelphia (CHOP).

Johnson himself helped design the recipe for the smoothies – called the Green Machine – which include kale, spinach, pineapple, honey, and banana. Johnson says he's a regular at Fuel, and this smoothie let him combine a bunch of his fave foods. The smoothies range in price from $4.19 to $6.19, and between now and the Super Bowl, $2 of every sale will go straight to CHOP.

"This smoothie is special to me because it contains all of my favorite flavors and ingredients and is Eagles' green," said Johnson in a press release. "This partnership lets me give back to one of my favorite local charities, CHOP. I have always loved kids and wanted to help out for those less fortunate, ill, or in need."

The smoothie is now on the menu at Fuel's Center City, University City, and East Passyunk locations. Between now and the Super Bowl, Fuel is donating proceeds; after the Super Bowl, the smoothie will remain on the menu for at least a week. And bonus: If the Eagles win the Super Bowl, Fuel will give away free small Green Machine smoothies from 11 a.m. to 1 p.m. on February 5.

Like what you're reading? Stay in touch with Be Well Philly—here's how: